Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering

Thibault Lesieur, Caterina de Bacco, Jess Banks, Florent Krzakala, Cris Moore, Lenka Zdeborová

To cite this version:
Thibault Lesieur, Caterina de Bacco, Jess Banks, Florent Krzakala, Cris Moore, et al.. Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering. 2016. cea-01448112

HAL Id: cea-01448112
https://hal-cea.archives-ouvertes.fr/cea-01448112
Preprint submitted on 27 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—We consider the problem of Gaussian mixture clustering in the high-dimensional limit where the data consists of $m$ points in $n$ dimensions, $n,m \to \infty$ and $\alpha = m/n$ stays finite. Using exact but non-rigorous methods from statistical physics, we determine the critical value of $\alpha$ and the distance between the clusters at which it becomes information-theoretically possible to reconstruct the membership into clusters better than chance. We also determine the accuracy achievable by the Bayes-optimal estimation algorithm. In particular, we find that when the number of clusters is sufficiently large, $r > 4 + 2\sqrt{3}$, there is a gap between the threshold for information-theoretically optimal performance and the threshold at which known algorithms succeed.

Clustering $m$ points in $n$-dimensional space is a ubiquitous problem in statistical inference and data science. It is especially challenging in the (practically relevant) context of high-dimensional statistics when the dimension $n$ is large (so that there are many degrees of freedom) but the number of points $m$ is only linear in $n$ (so that the available information is limited). An important special case is the problem of clustering points generated by Gaussian mixture models (GMM). These are probabilistic models where all the data points are generated from a mixture of a finite number $r$ of Gaussian distributions.

In this paper, we are interested in understanding the fundamental limits on our ability to cluster points generated from GMM in the high-dimensional regime, when $m,n \to \infty$ while $m/n = \alpha$ is finite. We consider both a computational and an information theoretic viewpoint, and wish to answer the following questions: Given data generated by a GMM, (i) what is the best possible estimate, information-theoretically, of the parameters of the Gaussian mixtures, and of the individual assignments of the points? (ii) Are such optimal estimates computationally feasible in practice by polynomial-time algorithms? (iii) How do standard, widely used, methods such as principal component analysis [1,2] compare with these optimal predictions?

We address these questions by taking advantage of the recent burst of activity in the related problem of low-rank matrix factorization, using the cavity method from statistical physics [3] and the associated approximate message passing (AMP) algorithm [4,5]. We shall not attempt at mathematical rigor here, but it is worth noting that recent progress on closely related problems [6-8] is likely transferable to the present situation.

I. MODEL AND SETTING

Consider data generated by a GMM with $r$ clusters: for each $k \in \{1,\ldots,r\}$ we draw each coordinate of $V_k \in \mathbb{R}^n$ from a Gaussian of zero mean and unit variance. We then generate $m = \alpha n$ points $x_1,\ldots,x_m \in \mathbb{R}^n$ independently as follows: for each $j$, choose $t_j \in \{1,\ldots,r\}$ uniformly at random, and set $x_j = \sqrt{n} \frac{\rho}{n} V_{t_j} + U_j$ where $U_j \in \mathbb{R}^n$ and has mean 0 and variance $\Delta$ in each coordinate. Here $\rho$ is a parameter playing the role of the signal to noise ratio and the $\sqrt{\rho/n}$ factor ensures that the displacement of the centers is of the same order as the fluctuation of a high-dimensional Gaussian point around the surface of a sphere of radius $n$.

We consider the limit $n \to \infty$, while $\alpha, \rho, \Delta = O(1)$.

Using this generative model makes the clustering problem a non-trivial one: the clusters overlap significantly; assigning points to clusters without errors is impossible and it is also impossible to recover exactly the parameters of the GMM. We thus ask instead what is the best possible overlap with the ground-truth clusters. It is instrumental to view this problem as noisy low-rank matrix factorization. Specifically, let $V$ denote the $n \times r$ matrix whose $k$th column is $V_k$, and let $S$ be the $m \times r$ matrix where $S_{jk} = 1$ if $t_j = k$ and 0 otherwise. We denote by $v_i, s_j \in \mathbb{R}^n$ the $i$th and $j$th rows of $V$ and $S$; note that $v_i$ is the vector of the $i$th coordinate of the cluster centroids. Then the observed data matrix $X$ is

$$X = \sqrt{\frac{\rho}{n}} VS^\top + U,$$ (1)

where $U$ is a Gaussian i.i.d. $n \times m$ random matrix with mean 0 and variance $\Delta$. Thus $X$ is a noisy observation of a matrix $VS^\top$ of rank $r$. We shall use without loss of generality $\Delta = 1$, as other values of $\Delta$ simply cause a rescaling of the signal-to-noise ratio $\rho$.

Assuming the data is generated by model (1), we want to know how well we can reconstruct the true cluster assignments $\{t_j\}$ and cluster centers $\{V_k\}$. To answer these questions, we use Bayesian inference. The generative model for the data is

$$P(X,S,V) = P(S)P(V)P(X|S,V)$$
$$= \prod_{i=1}^n P(v_i) \prod_{j=1}^m P(s_j) \prod_{i=1}^n \prod_{j=1}^m e^{-\frac{1}{2\rho}(X_{ij} - \sqrt{n} v_i^\top s_j)^2},$$
where \( P_v \) and \( P_s \) are the probability distributions from which the cluster centers \( v_i \) and assignments \( s_j \) are drawn. We take \( P_v \) to be Gaussian with mean zero and covariance matrix \( I_r \). Since the labels \( t_j \) are uniform, \( P_s \) is the uniform distribution over the \( r \) canonical basis vectors

\[
P_v(v) = \frac{1}{(2\pi)^{r/2}} \exp\left(-\frac{\|v\|^2}{2}\right),
\]

\[
P_s(s) = \frac{1}{r} \sum_{i=1}^{r} \delta(s - e_i).
\]

Using Bayes’ rule to compute the posterior probability of \( S \) and \( V \) given the data \( X \), one gets

\[
P(S, V | X) = \frac{1}{P(X)} \prod_{i=1}^{n} P_v(v_i) \prod_{j=1}^{m} P_s(s_j) \exp\left(-\frac{1}{2} \sum_{ij} (x_{ij} - \frac{v_i^{T} \cdot e_j}{\sqrt{\alpha}})^2\right).
\]

We denote the ground truth cluster centers and assignments used to generate \( X \) as \( V_0 \) and \( S_0 \). We assume that the correct prior distributions \( (2) \) and \( (3) \) are known to the algorithm and analysis; specifically, the correct number of clusters \( r \) and the parameters \( \rho, \Delta \). We call this the Bayes-optimal setting. The question of what happens when the model is incorrect is also interesting, but in this article we let it aside. We note, however, that, taking the limit \( \rho \to \infty \) and \( \Delta \to 0 \) in the Bayesian setting would correspond to the minimization problem that the \( k \)-means algorithm [9] is trying to solve. The objective of soft \( k \)-means corresponds instead to the Bayesian setting in the limit \( \rho \to 0 \) while retaining a nonzero noise level \( \Delta \). Spectral methods, on the other hand, are equivalent to Bayesian inference where one abandons the hard constraint that the labels \( t_j \) are discrete, and effectively replace \( P_s \) with a Gaussian prior.

The above Bayesian approach converts the problem of estimating an assignment of points to clusters that maximizes the number of correctly assigned points to the problem of computing marginal probabilities of the posterior distribution. The mean squared error in estimating \( V_0 \) from \( X \) is minimized by using the (marginalized) conditional expectation of \( V \) given the data [10]. For estimation of \( S_0 \) we aim to minimize the number of mis-classifications. The corresponding Bayes-optimal estimator \( \hat{S}^{\text{MaxProb}} \) is constructed by assigning each data point to the cluster that, according to the marginals of \( (4) \), it is most likely to belong to. Given \( \hat{S}^{\text{MaxProb}} \), the overlap between the estimated cluster assignments and the ground truth is defined as

\[
\text{ErrorRate} = \frac{1}{m} \text{Tr} \left( \hat{S}^{\text{MaxProb}} S_0 \right)
\]

\[
\text{Overlap} = \frac{\text{ErrorRate} - 1/r}{1 - 1/r}.
\]

The ErrorRate is the percentage of correctly assigned points. This rate is 1 when the reconstruction is perfect. When taking the assignment at random the ErrorRate is \( 1/r \). The normalization in (6) is introduced in order to have a quantity that varies between 0 meaning no better reconstruction that chance, to 1 meaning perfect reconstruction.

The marginal posterior probabilities of the \( v_i \) and \( s_j \) are, however, difficult to compute in general. This is because the interaction terms involving the observations \( X_{ij} \) couple these variables together. This is where we turn to techniques from the statistical physics of disordered systems which allow marginalization of joint distributions such as \( (4) \).

II. RELATION TO PREVIOUS WORKS

Study of the Bayes-optimal estimation in Gaussian mixture clustering in the regime of finite \( \alpha \) is restricted to the statistical physics literature. Results is statistic are either for \( \alpha \to \infty \) or for algorithms (such as spectral ones) that are suboptimal in the present setting. Existing literature, that treats the Bayes-optimal setting either considered the problem only algorithmically [5] or analyzes it only for two clusters, \( r = 2 \), [11–14]. The main contribution of this paper is to extend the analysis to the general (finite) number of clusters and realize that the case with several clusters behaves considerably differently from \( r = 2 \).

Our approach relies on Approximate Message Passing (AMP) algorithm and its theoretical analysis, a large part of which was born in the statistical physics, along with the cavity and replica approach [3, 15, 16]. In the present case, the AMP algorithm is very close to the TAP equations [17]. AMP for Gaussian mixture clustering is a case of AMP for low-rank matrix factorization as has been written in [4, 5]. State evolution (SE) [6, 18, 19] is a theoretical technique, closely related to the cavity method [15], that allows to exactly characterize the behavior of the AMP algorithm in the limit of infinite system size. A number of works on low-rank matrix factorization used recently this approach [20–22]. The AMP, the state evolution and associated Bethe free energy were presented recently in a rather generic form in [22] and we use these results extensively here in the specific case of GMM clustering and analyze the associated phase transitions and phase diagrams.

Proving the results presented here in full generality, for a general case of low-rank matrix factorizations, is a task that has attracted a lot of attention recently. In particular the replica prediction in the symmetric rank-1 case has been proven rigorously in a number of situations [7, 20, 21, 23] and a very generic proof now exists [8]. This strongly strengthens the claim that the replica predictions for the non-symmetric matrices (which we use here) are exact, despite the conjecture being still open.

III. SUMMARY OF MAIN RESULTS

By analyzing the Bayes-optimal estimation in the Gaussian mixture clustering we find that depending on the values of \( \rho \), \( r \) and \( \alpha \) (without loss of generality we consider from now on \( \Delta = 1 \)), the problem appears in one of the following three phases:

- **EASY:** The theoretically optimal reconstruction is better than chance both for the clustering problem and for finding the centroids and the AMP algorithm described in this article is able to reach this performance in the limit of large systems.
• IMPOSSIBLE: The theoretically optimal reconstruction does not perform better than chance. The matrix $X$ contains no exploitable information on the assignment of each point. For the centroids the best estimate is given by taking the mean of all the data points. No algorithm can exist to recover the assignment of the points better than chance.

• HARD: The theoretically optimal reconstruction gives a result that is better than chance both for the clustering problem and for finding the centroids. However, without using prior knowledge on the assignment of the points or on the centroids, the AMP algorithm is not able to converge toward its optimal fixed point. In fact, the AMP fixed point reached from an uninformed initialization does not provide any information on the assignment of the points. From the current knowledge, it is plausible that this phase is computationally hard not only for AMP but for all polynomial algorithms.

Depending on the number of clusters (that is, the rank $r$) and the sample-to-dimension ratio $\alpha$ we identified a boundary between two regimes at

$$r_c = 4 + 2\sqrt{\alpha}.$$  \hspace{1cm} (7)

- If the number of clusters is small enough, $r < r_c$ there are only two phases in the problem as the signal-to-noise parameter $\rho$ increases: The impossible and the easy one. This among others means that in the large size limit we are always able to reach Bayesian optimal reconstruction performance using the AMP algorithm. These two phases are separated by a sharp impossible/easy phase transition when

$$\rho_c = \frac{r}{\sqrt{\alpha}}.$$  \hspace{1cm} (8)

The problem is easy when $\rho > \rho_c$ and impossible otherwise.

- If instead the number of clusters is large enough, $r > r_c$ the three phases above are observed when decreasing $\rho$.
  - When $\rho > \rho_c = r/\sqrt{\alpha}$ we are in the EASY phase.
  - When $\rho_c > \rho > \rho_{\text{IT}}$ we are in the HARD phase.
  - When $\rho_{\text{IT}} > \rho$ we are in the IMPOSSIBLE phase.

IT here stands for information theoretic. We also compute the asymptotic behavior of $\rho_{\text{IT}}$ for a large number of clusters and found that

$$\rho_{\text{IT}}(r, \alpha) = 2 \sqrt{\frac{r \log r}{\alpha}} (1 + o_r(1)).$$  \hspace{1cm} (9)

This means that for $r$ large the HARD phase is very broad. Note that this last asymptotic result has been recently proven rigorously in [24] using the first and second moment methods, another indication of the correctness of the cavity/replica assumptions.

In what follows, we give statistical physics justifications for these claims, based on the analysis of AMP and on the interpretation of the Bethe free energy (related to the mutual information) as the exact one. We hope this will motivate further rigorous studies in this direction.

IV. APPROXIMATE MESSAGE PASSING (AMP)

We recall the AMP algorithm here using the notations of [22]. Let us define two denoising functions $f_v(A, B) \in \mathbb{R}^{r \times 1}$ and $f_s(A, B) \in \mathbb{R}^{r \times 1}$ as the means of the probability distributions

$$\frac{1}{Z_v(A, B)} P_v(v) \exp \left( B^T v - v^T A v \right), \hspace{1cm} (10)$$

$$\frac{1}{Z_s(A, B)} P_s(s) \exp \left( B^T s - s^T A s \right), \hspace{1cm} (11)$$

where $Z_v(A, B)$ and $Z_s(A, B)$ are normalization factors. Here $B$ is an $r \times r$ dimensional column vector and $A$ is an $r \times r$ matrix. For GMM clustering priors $\mathbb{E}$, we thus find

$$f_v(A, B) = (I_r + A)^{-1} B, \hspace{1cm} (12)$$

$$f_s(k, B) = \frac{\exp(B_k - A_{kk}/2)}{\sum_{l=1, \ldots, r} \exp(B_l - A_{ll}/2)}. \hspace{1cm} (13)$$

Taking the derivative of $f_v(A, B)$ and $f_s(A, B)$ with respect to $B$ yields the covariance matrices of (10) and (11), and we thus define

$$f'_v(A, B) = \left( \frac{\partial f_v}{\partial B} \right) \in \mathbb{R}^{r \times r}, \hspace{1cm} (14)$$

$$f'_s(A, B) = \left( \frac{\partial f_s}{\partial B} \right) \in \mathbb{R}^{r \times r}. \hspace{1cm} (15)$$

**Algorithm 1 AMP for clustering mixtures of Gaussians**

**Input:** Data $X$ and initial condition $\hat{\mathbf{v}}_{\text{init}}, \hat{\mathbf{s}}_{\text{init}}$.

For $t \geq 1$ compute

$$v_t, s_t \leftarrow \hat{\mathbf{v}}_{\text{init}}, \hat{\mathbf{s}}_{\text{init}}, \sigma_{v, i}^t \leftarrow 0 \mathbb{R}^{r \times r}, \sigma_{s, j}^t \leftarrow 0 \mathbb{R}^{r \times r}$$

$$\forall i \in [1, n], A_{v, i}^t \leftarrow \frac{\sqrt{n}}{\alpha} \sum_{j=1, \ldots, m} X_{ij} \sigma_{v, i}^{-1} - \frac{\alpha}{m} \sum_{j=1, \ldots, m} \sigma_{s, j} \sigma_{v, i}^t \sigma_{v, i}^{-1}$$

$$\forall i \in [1, n], \hat{\mathbf{v}}_{\text{init}} \leftarrow f_v(A_{v, i}^t, B_{v, i}^t), \sigma_{v, i}^t \leftarrow f'_v(A_{v, i}^t, B_{v, i}^t)$$

$$\forall j \in [1, m], A_{s, j}^t \leftarrow \frac{\sqrt{m}}{\alpha} \sum_{i=1, \ldots, n} X_{ij} \sigma_{s, j}^{-1} - \frac{\alpha}{n} \sum_{i=1, \ldots, n} \sigma_{s, j} \sigma_{s, j}^t$$

$$\forall j \in [1, m], \hat{\mathbf{s}}_{\text{init}} \leftarrow f_s(A_{s, j}^t, B_{s, j}^t), \sigma_{s, j}^t \leftarrow f'_s(A_{s, j}^t, B_{s, j}^t)$$

Here $\forall i \in [1, n] B_{v, i}^t \in \mathbb{R}^{r \times 1}, \forall j \in [1, m] B_{s, j}^t \in \mathbb{R}^{r \times 1}$ and $A_{v}^t \in \mathbb{R}^{r \times r}, A_{s}^t \in \mathbb{R}^{r \times r}$.

Iterate till convergence.

V. STATE EVOLUTION (SE)

The SE is a method for tracking the evolution of an AMP algorithm along iterations in the large system-size limit. A physics derivation of these SE equations can be found in
We first introduce the following order parameters
\[ M^t_v = \frac{1}{n} \sum_{i=1}^{n} \hat{v}_i^t \hat{v}_i^t \top, \]  
\[ M_s^t = \frac{1}{m} \sum_{j=1}^{m} s_j^{t} s_j^{t} \top, \]  
where \( \hat{v}_i^t \) and \( s_j^t \) are the estimators of the posterior means of the variables \( v_i \) and \( s_j \) at time \( t \) defined by
\[ \hat{v}_i^t = f_v(A^t_i, B^t_v, i), \]  
\[ s_j^{t+1} = f_s(A^t_j, B^t_s), \]
where the \( r \times r \) matrices \( M \) describe the overlap of the estimators with the ground truth solution. The SE reads [22]
\[ M^t_v = E_{W,v_0} \left[ f_v(\alpha \rho M^t_v, \alpha \rho M_s^t v_0 + \sqrt{\alpha \rho M^t_v W}) v_0 \right], \]  
\[ M_s^{t+1} = E_{W,s_0} \left[ f_s(\rho M^t_v, \rho M_s^t s_0 + \sqrt{\rho M^t_v} W) s_0 \right], \]
where \( W \) is an \( r \)-dimensional Gaussian variable with zero mean and unit variance in each coordinate, and \( v_0 \) and \( s_0 \) are random variables distributed according to \( P_v \) and \( P_s \) respectively.

A fixed point of the SE equations is a local extremum of the so-called Bethe free energy \( \phi_B \) [22]
\[ \phi_B(M_v, M_s) = \frac{\alpha \rho \text{Tr}(M_v M_s^\top)}{2} - E_{W,v_0} \left[ \log Z_v(\alpha \rho M_s \alpha \rho M_s v_0 + \sqrt{\alpha \rho M_s W}) \right] - \alpha E_{W,s_0} \left[ \log Z_s(\rho M_v \rho M_s s_0 + \sqrt{\rho M_v} W) \right], \]
where the quantities \( Z_v \) and \( Z_s \) are the normalization factors from (10) and (11). A stable fixed point of the SE equations above is a local minimum of \( \phi_B(M_v, M_s) \) as can be seen by taking partial derivatives with respect to \( M_v \) and \( M_s \).

It follows from the replica theory of statistical physics that if there are multiple stable fixed points of (20-21), only the one with minimal free energy \( \phi_B \) corresponds to the performance of the Bayes-optimal estimation. On the other hand the performance reached by the AMP algorithm corresponds to the fixed point of SE with largest error.

VI. ANALYSIS FOR MIXTURE OF GAUSSIANS

We analyze now the state evolution with the \( f_v \) and \( f_s \) functions corresponding to the GMM (12-13) for general number of clusters \( r \). The difficulty here lies in the fact that there is in general no analytic expression for updating equations (20-21). We observe that the following form of the order parameters is conserved under the update of SE equations for the \( \phi \)
\[ M^t_v = \frac{I_v b_s^t}{r} + (1 - b_v^t) J_r \]  
\[ M_s^t = b_v^t I_v + b_s^t J_r \]  
where \( (b_v^t, b_s^t, b_v^{t+1}) \in [0, 1]^3 \), and \( I_v \) and \( J_r \) are respectively the identity matrix and the \( r \times r \) matrix filled with 1. Having \( (b_v^t, b_s^t, b_v^{t+1}) = (1, 1, 0) \) would mean that we have achieved perfect reconstruction of the ground truth, while \( (b_v^t, b_s^t) = (0, 0) \) means that we are not able to extract any information from the matrix \( X \) beyond the average of the \( k \) clusters \( V_k \).

Rewriting eqs. (20-21) using (12-13) and (23-23) we get the following SE equations for the GMM
\[ b_v^t = b_v^t \rho + b_s^t \]  
\[ b_v^{t+1} = M_v \left( b_v^t \rho \right), \]
where
\[ M_v(x) = \frac{1}{r-1} \left\{ \frac{\int e^{\rho b_s^t \alpha - \rho b_v^t} \prod_{i=1}^{r} \mathcal{D} u_i - 1}{e^{\rho b_s^t \alpha - \rho b_v^t}} \right\}. \]
Where the \( u_i \) are Gaussian variables of mean 0 and unit variance. The equations close on \( b_v \) and \( b_s \) do not depend on the term \( b_v, J \). We can combine these to obtain a single update equation for the scalar variable \( b_v^t \)
\[ b_v^{t+1} = M_v \left( b_v^t \rho^2 \right). \]

The function \( M_v \), including the expansions for small \( \rho \) and for large \( r \), has been studied previously in [22]. From a numerical point of view \( M_v(x) \) can be effectively computed in the Monte Carlo scheme if one takes advantage of the permutation symmetry of the Gaussian variables.

A. Analysis of the phase transitions

Note that \( b_v = M_v(0) = 0 \) is always a fixed point that we will call uninformative. By expanding \( M_v(x) \) around 0 one gets:
\[ b_v^{t+1} = \frac{\alpha b_v^t \rho^2}{r} + \frac{\alpha^2 b_v^t \rho^4}{2} \left( r - 4 - \frac{2r}{\rho} \right) . \]

We are interested in when the uninformative fixed point becomes numerically unstable. From the expansion of (28) one deduces that this occurs when
\[ \rho > \rho_c = \frac{r}{\sqrt{\alpha}} . \]

Looking at the second derivative of (28), we deduce that when the uninformative fixed point becomes unstable, the second derivative is proportional to \( r - 4 - 2\sqrt{\alpha} \); if this is negative then this means that another stable fixed point appears close to 0 for \( \rho > \rho_c \). If on the other hand the second derivative is positive when \( \rho \) increases and crosses \( \rho_c \), then the new stable fixed point will not be close to zero and we see a discontinuous jump in the MSE achieved by initializing the iteration close to the uninformative fixed point. This phenomenon is known as a first-order transition. If one fixe the number \( r \), then we have a first order transition in the GMM if \( r > r_c = 4 + 2\sqrt{\alpha} \).

It turns out that this sufficient criteria is also necessary. In order to analyze the SE numerically we introduce two ways to initialize the equations
- Uninformative initialization \( b_v^t = 0 \) : Initializing in such a way is equivalent to assuming that we use
no knowledge about the ground truth signal. From an AMP point of view this starts with means of messages close to 0. We denote \( b_{\text{AMP}} \) the reached fixed point.

- Informative initialization \( b_{\text{inf}}^x = 1 \): This means initializing in the ground truth configuration. From an AMP point of view this corresponds to start with messages equal to the solution. Of course, in reality the algorithm would not have access to the solution; but here we apply this initialization in order to track the corresponding fixed point. We denote \( b_{\text{inf}} \) the reached fixed point.

For \( r < 4 + 2 \sqrt{\alpha} \) iterations with these two initializations lead to the same fixed point and the situation looks qualitatively as in Fig. 2 for two clusters. For \( \rho < \rho_c \), we reach the uninformative fixed point \( b_{\text{inf}} = 0 \), whereas for \( \rho > \rho_c \) we reach a fixed point with positive overlap with the ground truth, \( b_{\text{inf}} > 0 \).

If \( r > 4 + 2 \sqrt{\alpha} \), then depending on the value of \( \rho \) four different situations happen:

- \( \rho > \rho_c \): Here \( b_{\text{AMP}} = b_{\text{inf}} > 0 \): The AMP algorithm is able to reach the information theoretically optimal reconstruction. This is called the EASY phase.

- \( \rho_c > \rho > \rho_{\text{IT}} \): Here \( b_{\text{inf}} > b_{\text{AMP}} = 0 \), i.e. there are two fixed points to the SE equations and \( b_{\text{inf}} \) leads to a lower free energy. The AMP algorithm starting from the uninformative initialization is not able to reach the information theoretically optimal performance. Nevertheless it is possible, although in exponential time, to find a better fixed point with lower free energy. This is called the HARD phase.

- \( \rho_{\text{IT}} > \rho > \rho_{\text{sp}} \): Here still \( b_{\text{inf}} > b_{\text{AMP}} = 0 \), but now \( b_{\text{AMP}} \) has a lower free energy. The AMP algorithm starting from the informative state is able to reach a good fixed point correlated with the solution. However, finding this fixed point prior without knowledge of the solution is information theoretically impossible: it is hidden among an exponential number of other fixed points of AMP that have similar likelihood. This is called the information-theoretically IMPOSSIBLE phase.

- \( \rho_{\text{sp}} > \rho \): Here \( b_{\text{AMP}} = b_{\text{inf}} = 0 \), i.e. there is only one fixed point to the SE equations. This is also the information-theoretically IMPOSSIBLE phase.

To compute these transitions \( \rho_{\text{IT}} \) and \( \rho_{\text{sp}} \) numerically we consider one value of \( b_{\text{inf}} = M_r(x) \) and ask what is the value of \( \rho \) such that \( b_{\text{inf}} = M_r(x) \) is a fixed point. Using (27) the answer is

\[
\rho(x, r) = \frac{x}{2r} + \sqrt{\frac{x^2}{4r^2} + \frac{x}{\alpha M_r(x)}}.
\tag{30}
\]

The spinodal transition \( \rho_{\text{sp}} \) is the minimum value of \( \rho \) for which a fixed point other than 0 exists. We can estimate this by minimizing \( \rho(x, r) \) (30) with respect to \( x \) and at fixed \( r \). The information-theoretic transition \( \rho_{\text{IT}} \) is obtained by expressing the difference in the free energy between \( b_{\text{inf}} = 0 \) and \( M_r(x) \) at a given \( \rho \) and then requiring this quantity to be 0. It is possible to express the Bethe free energy using \( M_r \). If one integrates the gradient of \( \phi_B \) along the path \( g(u) \) defined by

\[
\forall u \in [0, M_r(x)], g(u) = \left( u, \frac{\rho(x)^2}{r} \rho \right)^{-1} \frac{1 + \rho(x)\rho(x)}{r}, \tag{31}
\]

after integrating by parts

\[
\phi_B(0, \rho(x, \alpha, r), \alpha; r) - \phi_B(M_r(x), \rho(x, \alpha, r), \alpha, r) = \frac{\alpha r - 1}{2r^2} \int_0^x du M_r(u) + \int_0^x du \frac{\rho(x)^2}{r} \rho - xM_r(x).
\tag{32}
\]

The information-theoretic transition is found where \( b_{\text{AMP}}(r, \alpha, \rho) \) and \( b_{\text{inf}}(r, \alpha, \rho) \) both have the same free energy. The behavior of \( \rho_c, \rho_{\text{IT}} \) and \( \rho_{\text{sp}} \) as a function of \( \alpha \) is illustrated in Fig. 1.

B. Large number of clusters

Formulas (30,32) also allow us to explore the large \( r \) limit of these solutions. From [22] we know that

\[
\forall \beta > 0, \lim_{r \to \infty} M_r(\beta r \log r) = 1_{\beta > 2}.
\tag{33}
\]

We can compute the asymptote of \( \rho_{\text{IT}} \) and \( \rho_{\text{sp}} \) when \( r \to \infty \). We can do that by setting \( x = \beta r \log(r) \) and then replacing \( M_r \) by (33) in (30) and (32).

To get \( \rho_{\text{sp}} \) one minimizes (30) with respect to \( \beta \). After some computation one gets

\[
\lim_{r \to \infty} \frac{\rho(\beta r \log(r), \alpha)}{r \log(r)} = \begin{cases} +\infty, & \text{if } \beta < 2 \\ \sqrt{\beta/\alpha}, & \text{if } \beta \geq 2 \end{cases}
\tag{34}
\]

Therefore in the large \( r \) limit \( \rho(\beta r \log(r)) \) is minimized by taking \( \beta = 2 \) and one gets

\[
\rho_{\text{sp}}(r, \alpha) = \sqrt{\frac{2r \log r}{\alpha}} (1 + o_r(1)).
\tag{35}
\]
To get $\rho_{TT}$ one finds the $\beta$ such that \[32\] is set to 0. In order to do that we write $x_{TT} = \beta_{TT} r \log(r)$. We take $\beta_{TT} > 2$ since we want to have $\rho$ to be above $\rho_{\text{sep}}$ so that there are multiple fixed point to the SE equations. For $\beta > 2$ to leading order one has

$$\rho(x = \beta r \log(r), r) = \sqrt{\frac{\beta r \log(r)}{\alpha}}(1 + o_r(1)).$$

Therefore,

$$\rho(x = \beta r \log(r), r) \approx \frac{1}{\alpha}.$$ \[37\]

Using this, equation \[32\] can be further simplified. One gets

$$\phi_B(0, \rho(x_{TT}, \alpha, r), \alpha; r) - \phi_B(\mathcal{M}_r(x_{TT}), \rho(x_{TT}, \alpha, r), \alpha, r)$$

$$= 0 \approx \alpha \int_0^{x_{TT}} du \mathcal{M}_r(u) - x_{TT} \mathcal{M}_r(x_{TT}) - \frac{1}{2(\beta_{TT} > 2)}.$$ \[38\]

By setting $x_{TT} = \beta_{TT} r \log(r)$ in this equation and taking the $r \to \infty$ limit one gets

$$0 = \int_0^{\beta_{TT}} du \mathcal{1}(u > 2) - \frac{\beta_{TT} \mathcal{1}(\beta_{TT} > 2)}{2}.$$ \[39\]

This is solved for $\beta_{TT} = 4$, therefore one has

$$\rho_{TT}(r, \alpha) = \frac{1}{2} \sqrt{\frac{r \log(r)}{\alpha}}(1 + o_r(1)).$$ \[40\]

Thus, for large number of clusters $r$, the gap between the information-theoretic detectability threshold $\rho_{IT}$ and the algorithmic threshold $\rho_c = r/\sqrt{\alpha}$ becomes large.

C. Algorithmic comparison

Fig. 2 and 3 contain numerical experiments with the AMP algorithm, and comparison with the theoretically predicted performance given by state evolution analysis. Data for both $r = 2$ and $r = 20$ clusters are presented. These two cases have qualitatively different properties. As predicted, the case with $r = 20$ clusters exhibits a first order transition: there is a sharp jump in the overlap at $\rho_c$. Instead the case $r = 2$ exhibits a second order transition: the overlap is continuous, only its derivative has a discontinuity.

In both Fig. 2 and 3 we also compare to the performance of the principal component analysis (PCA) performed on the matrix $X$. PCA is a standard spectral method to solve data clustering, one computes $r$ leading singular vectors of $X$ and instead of clustering $m$ points in $n$ dimensions, one concatenates the singular vectors into $m$ $r$-dimensional vectors and clusters in the $r$-dimensional space which is much simpler. The overlap reached with the PCA clustering follows from a more general theory of low-rank perturbations of random matrices [2], but it can also be derived from the state evolution analysis of AMP as we present in the appendix. In this case of GMM with equal-size clusters the phase transition observed in PCA coincides with the phase transition of AMP $\rho_c$. Concerning the performance as measured by the overlap in \[6\], we observe that although for two clusters the difference between the performance of the sub-optimal PCA and the Bayes-optimal AMP is hardly visible, for 20 clusters the performance of AMP close to the algorithmic transition $\rho_c$ is considerably better.

VIII. Conclusion

We analysed the problem of clustering high-dimensional data generated by the Gaussian mixture model. We computed the asymptotic accuracy of the Bayes-optimal estimator and compared it to the accuracy obtained by the approximate message-passing algorithm. We located phase transitions in both the information-theoretic and algorithmic performance.

Our main result is that, when the number of clusters is sufficiently large, eq. \[7\], there is a gap between the information-theoretic threshold—where it becomes possible to label points and find the cluster centers better than chance—and the computational threshold, at which polynomial-time algorithms such as PCA or AMP succeed. This suggests that, as has been conjectured for
Let us define the top eigenvectors of \( \text{PCA on the Gaussian mixture clustering problem.} \) Let us write the density evolution equation the posterior measure (4) and where we have replaced the Bayes-optimal prior on \( v_i \) and \( s_j \) with Gaussian priors

\[
P_v(v_i) = \frac{\exp\left(-\frac{||v_i||^2}{2}\right)}{\sqrt{2\pi}}, \quad P_s(s_j) = \frac{\exp\left(-\frac{r||s_j||^2}{2}\right)}{\sqrt{2\pi r^r}}.
\] (41)

Let us define \( V_{\text{PCA}}, S_{\text{PCA}} \) the matrices containing the first \( r \) singular vectors of \( X \). Where \( V_{\text{PCA}} \in \mathbb{R}^{n \times r}, S_{\text{PCA}} \in \mathbb{R}^{m \times r} \). \( S_{\text{PCA}} \) defines \( m \) points in a \( r \) dimensional space. We expect the \( r \) dimensional points extracted from \( S_{\text{PCA}} \) to form a mixture of Gaussians. Let us write the AMP for Gaussian priors. The general AMP equations for the mismatching priors can be found in [5]. One gets

\[
\hat{V}^t = \left( \frac{\rho}{n} X S^t - \rho \alpha \hat{V}^{t-1} \Sigma^t_v \right) \Sigma^t_v,
\] (42)

\[
\Sigma^t_v = (I_r + \rho/n \hat{S}^T S^t)^{-1},
\] (43)

\[
\hat{S}^{t+1} = \left( \frac{\rho}{n} X^T \hat{V}^t - \rho \alpha \hat{S}^t \Sigma^t_v \right) \Sigma^{t+1}_s,
\] (44)

\[
\Sigma^{t+1}_s = (rI_r + \rho/n \hat{V}^t V^t)^{-1}.
\] (45)

We aim to prove that at a fixed point \( \hat{V} \) and \( \hat{S} \) will be singular values of \( X \).

After simplifications of the above equations one gets

\[
\hat{V} = \left( \frac{\rho}{n} X \hat{S} \left( \Sigma_v^{-1} + \rho \alpha \Sigma_s \right)^{-1} \right),
\] (46)

\[
\hat{S} = \left( \frac{\rho}{n} X^T \hat{V} \left( \Sigma_s^{-1} + \rho \Sigma_v \right)^{-1} \right).
\] (47)

By putting this in (46) in (47) one gets

\[
\hat{S} = \frac{\rho}{n} X^T \hat{S} \left( \Sigma_v^{-1} + \rho \alpha \Sigma_s \right)^{-1} \left( \Sigma_s^{-1} + \rho \Sigma_v \right)^{-1}.
\] (48)

This means that the columns of \( \hat{S} \) are a linear combination of eigenvectors of \( X^T X \) that is right eigenvector of \( \hat{S} \).

We can also write the corresponding state evolution equations. One gets

\[
M^t_v = \mathbb{E}_{W,v_0} \left[ f_v \left( \alpha \rho M^t_s, \alpha \rho M^t_s v_0 + \sqrt{\alpha \rho M^t_s} W \right) v_0 \right],
\]

\[
Q^t_s = \mathbb{E}_{W,v_0} \left[ f_v \left( \alpha \rho M^t_s, \alpha \rho M^t_s v_0 + \sqrt{\alpha \rho M^t_s} W \right) f_v(v_0, \cdots, \cdots) \right],
\]

\[
M^{t+1}_s = \mathbb{E}_{W,v_0} \left[ f_s \left( \rho M^t_s, \rho M^t_s v_0 + \sqrt{\alpha \rho M^t_s} W \right) v_0 \right],
\]

\[
Q^{t+1}_v = \mathbb{E}_{W,v_0} \left[ f_s \left( \rho M^t_s, \rho M^t_s v_0 + \sqrt{\alpha \rho M^t_s} W \right) f_s(v_0, \cdots, \cdots) \right],
\]

With the Gaussian priors (41) this gives us

\[
M^t_v = (I_r + \alpha \rho Q^t_v)^{-1} \alpha \rho M^t_s,
\] (49)

\[
Q^t_s = (I_r + \alpha \rho Q^t_v)^{-2} \left( \alpha^2 \rho^2 M^t_s + \alpha \rho Q^t_s \right),
\] (50)

\[
M^t_s = (rI_r + \alpha \rho Q^t_v)^{-1} \rho M^t_s/r,
\] (51)

\[
Q^t_s = (rI_r + \rho Q^t_v)^{-2} \left( \rho^2 M^t_s/r + \rho Q^t_s \right).
\] (52)

Experiments with these equations show that stable fixed point always are always of the form

\[
M^t_v = m^t_v R, \quad Q^t_s = m^t_s I_r,
\] (53)

\[
M^t_s = \frac{m^t_s R}{r}, \quad Q^t_s = \frac{m^t_s I_r}{r},
\] (54)

where \( R \) is some rotation matrix. The new update equation become

\[
m^{t+1}_v = \frac{\rho \alpha m^t_s}{1 + \rho \alpha m^t_s},
\] (55)

\[
m^{t+1}_s = \frac{\rho m^t_v}{r + \rho m^t_v}.
\] (56)

Let us write

\[
m^t_v = m^t_v \rightarrow \infty = \max \left( \frac{\alpha \rho^2 - r^2}{r + \rho \alpha}, 0 \right),
\] (57)

\[
m^t_s = m^t_s \rightarrow \infty = \max \left( \frac{\alpha \rho^2 - r^2}{r \rho \alpha + \rho^2 \alpha}, 0 \right),
\] (58)

By combining (45), (43), (53) and (54) one sees that in the large \( n \) limit \( \Sigma_v \) and \( \Sigma_s \) are proportional to the identity. Therefore using (48) one gets

\[
\hat{S} = \frac{\rho}{n} X^T X \hat{S},
\] (59)

where \( \alpha \) is some number. This means that \( \hat{S} \) is proportional to the first eigenvectors. But the AMP analysis also tells us how \( \hat{S} \) will be distributed. The details of the analysis of state evolution can be found in [22] in the \( XX^T \) case

\[
\hat{s}_j = f_s \left( \rho m_v I_r, \rho m_v R s_{0,j} + \sqrt{\rho m_v} W \right).
\] (60)

Where \( W \) is a Gaussian variable given by \( \mathcal{N}(0, I_r) \). Up to a proportionality constant one has

\[
\hat{s}_j = \sqrt{\rho m_v} R s_{0,j} + W.
\] (61)

This means that the first \( r \) right eigenvectors of \( R \) are distributed as a mixture of Gaussians in a \( r \) dimensional space where the centers are placed at positions \( \sqrt{\rho m_v} R e_k \) and the noise is a Gaussian white noise of zero mean and covariance matrix \( I_r \). The questions now becomes how well can one cluster this mixture of Gaussians? Since we are dealing with a finite-dimensional space \( r = O(1) \) and a large number of points \( m \rightarrow \infty \) then we know that we should be able to learn the parameters from this mixture of Gaussian perfectly.

The problem now becomes given a point \( v_j \) what is the chance that it was created using point \( v_0,j \) this is done by
maximizing the likelihood. To find back with what $e_k \hat{v}_j$ was created one needs to compute.

$$k(\hat{v}_j) = \arg\max \left\{ k \in [1; r], \frac{1}{r\sqrt{2\pi}} \exp \left( -\frac{1}{2} ||\hat{v}_j - \sqrt{\rho_m e_k}||^2 \right) \right\}$$

or

$$k(\hat{v}_j) = \arg\max \{ k \in [1; r], \langle \hat{v}_j; \hat{e}_k \rangle \} . \tag{62}$$

One need to compute with what probability $\hat{k}$ is the right $k$

$$P(\hat{k}(\sqrt{\rho_m e_k} + W) = k_0) = \max_{k \in [1; r] , k \neq k_0} \left\{ W_k \right\} .$$

This can be computed using the distribution of the maxima of random variables of zero mean and variance $r - 1$ Gaussian variables.

$$P(\hat{k}(\sqrt{\rho_m e_k} + W) = k_0) = \int_{-\infty}^{+\infty} du \left[ 1/2 - 1/2\text{erf} \left( u - \sqrt{\rho_m e_k} \right) \right] P_{\max}^{-1}(u), \tag{63}$$

where $P_{\max}^{-1}$ is the density probability of the maximum of $r - 1$ independent Gaussian variables of mean 0 and unit variance. Where $\text{erf}(x) \in [-1; 1]$ is defined as $\text{erf}(x) = \sqrt{\frac{2}{\pi}} \int_0^x \exp \left( -\frac{u^2}{2} \right) du$.

$$P_{\max}^{-1}(u) = \frac{r - 1}{\sqrt{2\pi}} \exp(-u^2/2) \left[ 1 + \frac{\text{erf}(u)}{2} \right]^{-r/2} . \tag{64}$$

Therefore one can compute the average number of errors made using PCA to cluster the data points.

$$\text{ErrorRate}_{PCA} = \int_{-\infty}^{+\infty} du \left[ 1 - \frac{1}{2} \text{erf} \left( u - \sqrt{\rho_m e_k} \right) \right] P_{\max}^{-1}(u).$$

The mean-square error achieved by PCA is then $r(\rho - \rho_v)$ and is known in the theory of low-rank perturbations of random matrices [2].

REFERENCES


